图书情报工作
第63卷 第7期 2019年4月

数字人文研究领域的知识图谱构建与分析*
——基于WoS文献关键词和引文上下文的实证

许鑫 陈路遥 杨佳颖
华东师范大学经济与管理学部信息管理系 上海 200241

摘要：[**目的/意义**]引文是施引文献与被引成果的纽带，反映了后继者的借鉴和肯定。本研究在传统题录关键词网络的基础上，创新地将引文上下文关键词作为研究材料，所构建知识图谱不仅能揭示文献主题的深层次信息，也能够反映受众主观筛选和利用文献的知识过程。[**方法/过程**]选取数字人文为研究领域，获取3个文献集和两个引文文本集，构建两个无向的关键词共现网络和两个有向的基于文献引证的关键词网络。通过共现网络，观察数字人文领域知识的吸收与扩散；通过引证关键词网络，观察数字人文的形成与转化。[**结果/结论**]研究揭示数字人文的研究重点、核心领域与核心技术，从受众的角度为数字人文领域未来研究提供借鉴和参考。

关键词：知识图谱 数字人文 引文上下文 关键词网络 可视化
分类号：G253
DOI：10.13266/j.issn.0252-3116.2019.07.011

1 引言

在图情领域，知识图谱被定义为显示科学知识发展进程与结构关系、有助于知识发现的领域知识地图，是将既定主题下的抽象科学信息映射入空间结构和图形的网状化可视化方法[1]。知识网络能够实现知识的创造与传递，特别是对于领域知识网络，能反映出一个领域内部知识的流动与传播。但在传统知识网络研究中，大多基于标题、摘要、关键词等具有作者主观性的信息所构建，反映了从创作者角度传递出的知识信息，无法反映出受众角度所获得的真正信息。对受众主动阅读、筛选、取舍、利用的引文文本进行研究，能够挖掘出在显性信息中难以发现的潜在知识，充分发挥出引用内容的价值和作用。

数字人文的前身是人文计算。人文计算侧重计算机学科在传统人文学科的应用[2]，但随着信息时代的到来和数字技术的普及，人文学者借助数字思维来解决人文问题的动机在逐渐增强，研究的落脚点逐渐从狭义的计算思维转向含义丰富的数字化，从组织方法创新转向人文内容本身，故而引申出数字人文的概念。数字人文作为新兴领域，概念内涵也在不断演变。知识图谱既是数字人文的实现手段之一，也是数字人文领域研究脉络梳理的重要方法之一。

本文以数字人文研究领域为对象，结合共现网络与引证网络，构建数字人文领域的知识网络，创新性地将引文文本中提取出的关键词作为网络节点，来构建引用关系网络。从显性信息深入到隐性信息，赋予原有引证网络更丰富的资料内容，有利于发现数字人文领域内的潜在关联。从知识扩散和知识吸收两个视角，梳理数字人文知识网络的演变路径，有助于把握数字人文领域未来的发展之路。

2 文献综述

科学知识图谱，即科学计量学的知识图谱概念[3]。在知识网络预测中，通过研究由分离集形成的引证网络、合作网络或二分网络，能够对知识单元的联系进行识别[4]。科学知识图谱被广泛应用于科学信息与知识的生产、展示与传播。

在本研究所切入的词层面上，当前学者主要是利用现有的作者所标注的文献关键词，进一步进行共现、

* 本文系上海市哲学社会科学规划项目"上海图书馆特藏资源的数字人文应用研究"（项目编号：2017BTQ001）研究成果之一。
作者简介：许鑫（ORCID：0000-0001-7020-3135），教授，博士，博士生导师，E-mail：xxu@infor.ecnu.edu.cn；陈路遥（ORCID：0000-0002-9740-8664），硕士研究生；杨佳颖（ORCID：0000-0003-1823-2785），硕士研究生。
收稿日期：2018-07-29 修回日期：2018-11-07 本文起止页码：86-95 本文责任编辑：徐健

聚类或耦合分析等,从而识别研究前沿与热点[5-6]。宋艳辉等利用作者关键词耦合的方法,与以文献为计量单位的常规方法进行比较,发现以关键词为计量单位能得到更为直观清晰的研究内容[7]。关键词除了与内容关联更密切,还能反映知识的吸收与扩散作用,张玲玲等通过提炼文献主题关键词,梳理了知识的扩散方向与脉络[8]。罗双玲等认为引文关键词的集聚形成了社区结构,并称高频标题关键词网络为"主题社区"[9]。

施引是受众阅读原始文献之后,主动发生的引用行为。引文是施引者归纳被引文献主题所形成的文本。引文上下文作为其延伸部分,能够揭示被引文献更深层次的内容,故而具有一定的研究价值[10]。引文上下文是后继研究者对被引成果的借鉴和肯定,是施引文献和被引文献通过认知建立起的关系。Y. Liu 等在知识融合和扩散的框架体系里,表明文章用词和引文的文章用词都反映了知识融合的影响作用[11]。目前利用引文上下文进行网络分析的研究较少,L. Bornmann 等以人为研究对象,对计量学家 E. Garfield 的相关引文的上下文本进行共现网络分析,发现引文上下文比施引文献的标题和摘要更能反映被引文献的内容[12]。

在开放式创新的大格局下,国内外网络绘制相关工具的大量兴起,促进了知识图谱的研究路线不断被重视与拓展,其中陈超美团队基于 JAVA 语言开发的 CiteSpace 软件在国内最负盛名[13]。利用该工具,肖明将 CiteSpace 相关文献关键词、机构、作者、期刊等基本科学信息构建网络图谱,然而仅仅停留在显性信息的分析上[14]。国内外利用知识网络方法,主要对不同的领域,尤其是新兴领域,进行研究脉络的梳理与热点探测[15-16],数字人文即是其中一个重要应用方向。

数字人文的发展离不开其所处的时代背景,并逐渐为人文学科研究带来了质的变革。S. Schreibman 等梳理归纳了数字人文早期的研究历史,覆盖了考古、美术、文学、音乐、表演、多媒体等领域[17]。数字人文的研究内容也在不断丰富,其内涵从人文计算延伸到数据存储、数据组织、可视化分析等数字技术相关的方方面面,如 D. Cooper 等提取并编码地名或空间相关的文字内容,使之以地图形式呈现[18];U. Hinrichs 等构建了科幻小说的物件、词云、符号、时间线的合集,并将其可视化[19]。图书馆人和信息科学学者作为数字人文的核心贡献者,在教学服务、凝聚用户、人才培养、资源建设方面都有所贡献[20]。

在数字人文研究成果梳理方面,并不乏采用科学知识图谱进行研究的文献。得益于 CiteSpace 等有效的工具,国内学者是该研究思路的主力军:主要通过文献的共现网络或引证关系归纳数字人文研究的热点领域,并剖析数字人文研究的演化路径[21-22],该方法在国际层面也逐渐得到认可[23]。然而,这类文章将一种关系作为研究重点;同时往往基于领域内的显性知识,并没有挖掘出受众角度的深层次隐性知识。

目前国内外数字人文研究均处于文献积累阶段,国内研究较国外起步晚,且成果量与国外存在一定差距。现有总结性文献在研究思路上,将共现网络与引证网络割裂开来,没有结合考虑两种关系;在研究方法上,当前研究仅仅关注创作者传递出的显性信息,而忽略了施引方通过阅读提炼得到的引用文本里的潜在信息。所以,数字人文领域正需要新的思路与方法,来对历史研究脉络进行梳理与把握。

3 研究方案设计

3.1 数据获取与定义

本研究将 Web of Science 核心合集作为主要数据来源,构建 3 个文献集(即中心文献集、参考文献集、引文文本集),通过获取全文来建立两个引文文本集(即中心文献的引文文本集、施引文献的引文文本集)。在获取全文过程中,各全文数据库与互联网搜索引擎提供的资源也同时被参考,以确保数据完全。

中心文献集,被定义为以"digital humanities"作为关键词进行主题字段检索所获得的 768 条文献数据(截止时间 2018 年 1 月)。施引文献集,被定义为中心文献对应的1 100条施引文献记录,可以通过点击 Web of Science 各文献的被引记录下载获取。参考文献集,被定义为被引次数最高的 20 篇中心文献的参考文献信息,共包括 956 条文献记录。

在获取 3 个文献集后,分别提取各文献所原始标注的关键词,并根据各集合的论文记录,进一步下载论文全文。

引文文本集包括中心文献的引文文本集(中心文献引用参考文献的引文文本集)和施引文献的引文文本集(施引文献引用中心文献的引文文本集)。研究通过参考文献列表与引用标识符信息,提取出相关文献的引文上下文文本。

3.2 引文上下文关键词的识别

引文上下文是引文内容分析的基础,对其的识别和相关应用已成为研究热点。在 A. Bader 对引文的研究中,选取了不同长度的引文窗口进行测试实验:在

引用标识的前后截取共 10 个单词、30 个单词、50 个单词以及不计单词个数的引用句(以引用标识所在的整句作为引用句,通过标点符号界定)共 4 组数据,结果证明,引文窗口长度为 50 的引文上下文最能代表被引文献的内容[24],与学者 S. Bradshaw[25]早前的研究结果相印证。因此,本文选取包含引用标识附近 50 个单词的引文上下文文本形成资料集。

具体来说,在提取引用标识附近 50 个单词作为引文文本集时,遵循以下规则:①考虑文献结构,引文上下文的截取必须在同一段落内。②一个引文上下文文本中只能包含一个引文标识。当一个引文句中包含多个引用标识时,除了引用标识同时出现或以"and"进行连接两种情况外,则需要缩短引文上下文的长度,通过在句子边界断句来实现。③对同一文献进行多次引用时,则保留多个引文上下文文本。以引用标识为唯一标识符,每出现一次引用标识,就截取一个引文上下文文本。

在提取出引文上下文文本之后,利用下述方法识别文本关键词:

首先,将所有引文上下文文本视为一个整体,进行 LDA(latent Dirichlet allocation)主题识别。LDA 是一种文档主题生成模型,也称为 3 层贝叶斯概率模型,包含词、主题和文档 3 层结构。通过调用 Python 的 sklearn 模块,设定忽略在 50% 的上下文文档语料库中都出现的常用词语(max_df = 0.5),识别出引文窗口前 10 个主题,以及每个主题的前 10 个关键词,以了解引文文本的主要研究主题和方向。其次,将所有引文文本视为一个整体,对所有引文文本进行切词与词频统计,利用中心文献集中的原有关键词构建自定义词表,通过人工标注的方式建立停用词表和同义替换词表。根据此 3 个词表,对原词频统计结果进行处理,转化为词的权重表;剔除停用词;合并同义词;提高自定义词权重。最后,以此权重词表为依据,再回到每一个引文文本中,至多提取出每一个引文文本中权重最高的 5 个词,成为引文上下文的关键词。

3.3 研究方案

知识网络有助于厘清文献间的关系,从文献中提取信息和知识单元,进一步了解知识的结构和演化。现有的词的共现网络研究主要反映研究主题及主题间的联系,以了解当前研究的热点与研究类群关系。

现有科学引证网络大多是以文献或作者为节点,以文献之间的引用关系作为节点之间的联系边,以此构建相关引用文献之间的引用网络。通过科学引证网络,可以了解科学知识的传播与流动,发现其中的传承发展或转化创新关系,也可以研究领域科学知识发展的脉络和结构。

关键词对文献主题具有更直观的揭示作用,是文献内容的浓缩。借鉴上述方法,构建关键词的引用关系网络,反映论文中一个关键词构建的文献情景对另一个关键词构建的文献情景的引用。本文将其定义为基于文献引证的关键词网络,简称为引证关键词网络。

具体地,从两个切入点来对数字人文领域进行研究:一是构建关键词共现网络,以了解知识的结构;二是构建基于引证的引文上下文关键词网络与引用文献关键词网络,以了解知识的脉络。

在关键词共现网络中,通过合并关键词集,构建数字人文核心文献关键词的知识吸收网络和知识扩散网络。合并,是指在保留词与词的共现关系基础上,对不同词集中的相同关键词进行连接。知识吸收网络,是合并参考文献集、中心文献的引文文本集和中心文献集的关键词共现网络。知识扩散网络,是合并中心文献集、施引文献的引文文本集和施引文献的关键词共现网络。研究将知识的动态演变转为了静态网络,以此对数字人文领域核心知识的吸收与扩散进行分析。

在基于引证的关键词网络中,则考虑了中心文献和施引文献的关系、高被引中心文献与参考文献的引用关系。在中心文献集与施引文献集的引用关系中,被引关键词是中心文献集文献原有关键词,施引关键词是施引文献引用中心文献的引文文本关键词;在高被引中心文献集与参考文献集的引用关系中,被引关键词是参考文献集文献原有关键词,施引关键词是中心文献引用参考文献的引文文本关键词。

基于引用关系,把被引关键词、施引关键词作为网络的节点,引用关系作为边,构建关键词引用网络。从中心文献与施引文献的关系中,可以发现数字人文研究的发展与转化趋势;从中心文献与参考文献的关系中,可以发现数字人文研究的来源与形成;通过对其来源与去向进行综合分析,可以综合了解数字人文的"前世今生",厘清数字人文发展的整个知识脉络。

4 基于核心文献关键词的数字人文知识共现网络

4.1 知识吸收网络

中心文献集中的关键词代表了当前数字人文领域中的核心知识,可以认为这些知识来源于对参考文献集中核心知识的引用转化。基于此,将参考文献集、中

心文献引文文本集和中心文献集的关键词以共现关系进行合并，可以综合观察数字人文核心文献关键词的知识吸收过程和结果。

图 1 是合并后的关键词共现网络，共有 2 968 个不同的关键词，构成了共计 13 069 对共现关系对。从图 1 中可以看到，数字人文知识的吸收与形成中，朝两团簇核心关键词聚拢，边缘游离着部分独立的小网络。

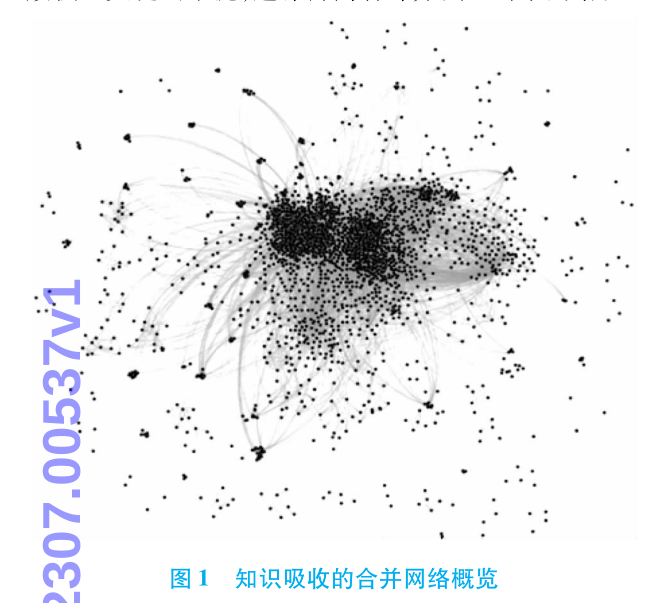


图 1　知识吸收的合并网络概览

知识吸收的合并网络的 h 强度为 48，即在网络中至少有 48 条联系的强度不低于 48。进一步，通过 h 强度精炼这一网络，得到如图 2 所示的数字人文核心文献关键词的知识吸收网络的 h 子网。即在图 2 中，包括了整体网中联系强度不低于 48 的联系，以及这些联系所连接的节点。

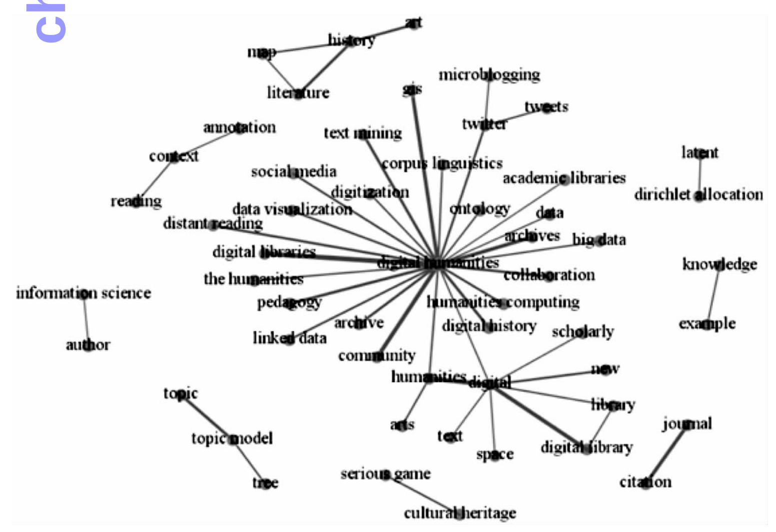


图 2　数字人文核心文献关键词的知识吸收网络的 h 子网

从图 2 中可以看到，主要形成了一个以数字人文概念为中心的核心网络，其余有一些独立概念所形成的小网络。在核心网络中，数字人文概念与人文学科、教育学、数字史学等学科关联性最强，表明这几类学科是数字人文关键知识吸收过程中的重要学科；与文本挖掘、数据可视化、关联数据等技术路线关联性最强，表明这几类技术在数字人文关键知识吸收过程中产生的影响力最大；除此之外，与人文计算、GIS、远距离阅读、数字图书馆、社交媒体、大数据等概念也极具关联性，表明这些概念相关知识影响着数字人文领域知识的吸收与形成过程。在核心网络边缘，有两个小分支，一个是 Twitter 与微博等形成的社交网络知识结构；另一个是数字化与文本、图书馆、创新性等形成的过程性知识结构。

除了核心网络，还存在着一些独立小网络：①文学、艺术、史学、地图所构成的小网络，反映了基于地图技术对文学、艺术等进行历史研究的跨学科性；②主题模型、主题和树所构成的小网络，反映了基于树的主题模型在数字人文领域前期是相对重要的一项技术方法；③文化遗产和严肃游戏，这一强关系在整个网络中具有较高的独立性，是数字人文研究中的一个小分支，致力于基于严肃游戏的模式展示和传播文化遗产；④期刊和引用的关系，主要是基于引用数据和指标对期刊进行评价。

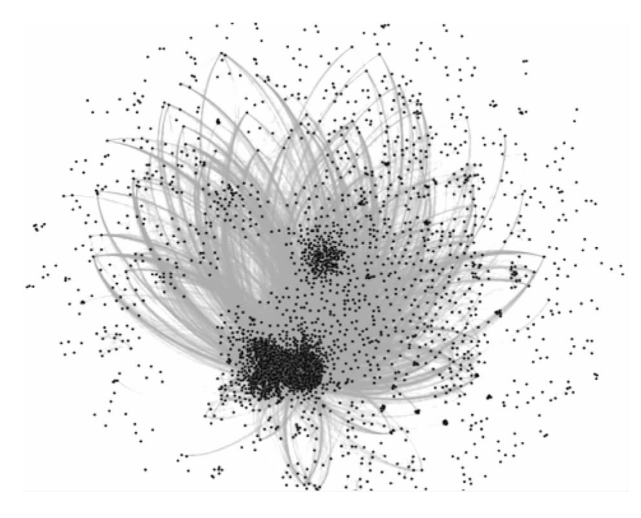
4.2　知识扩散网络

中心文献集的核心知识在传播过程中，对后续知识的形成产生了影响，具有知识扩散传播的过程。基于此，将中心文献集、施引文献引文文本集和施引文献集的关键词以共现关系进行合并，可以综合观察数字人文核心文献关键词的知识扩散过程和结果。

图 3 是合并后的关键词共现网络，共有 3 790 个不同的关键词，构成了 60 366 对共现关系对。从图 3 中可以看到，数字人文知识的扩散与传播中，主要是以一个关键词团簇为核心进行发散，其间还产生了一些小的知识网络。

知识扩散的合并网络的 h 强度为 42，即在网络中至少有 42 条联系的强度不低于 42。通过 h 强度精炼这一网络，得到如图 4 所示的数字人文核心文献关键词的知识吸收网络的 h 子网。即在图 4 中，包括了整体网中联系强度不低于 42 的联系，以及这些联系所连接的节点。

从图 4 中可以看到，数字人文核心知识在扩散中，


图 3　知识扩散合并网络概览

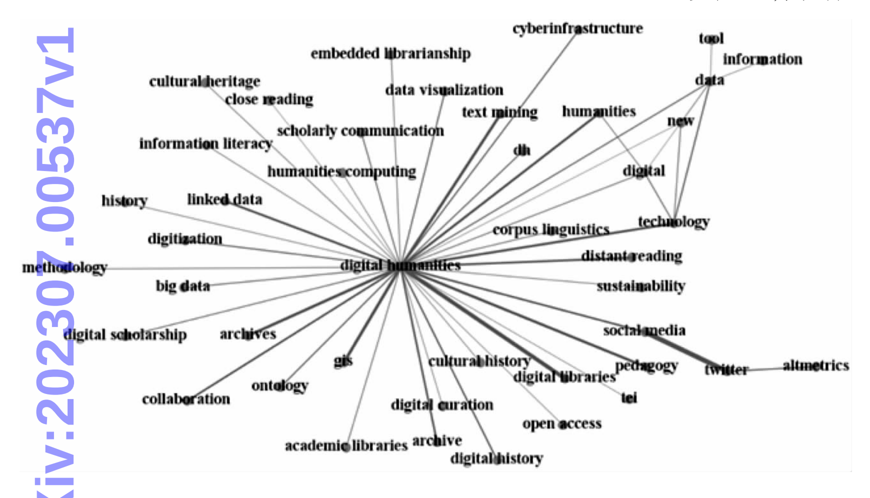


图 4　数字人文核心文献关键词的知识扩散网络的 h 子网

主要就围绕着数字人文的概念进行，整个 h 子网主要呈中心发散状。在其中，图书馆类知识表现特别突出，如嵌入式图书馆、数字图书馆、学术图书馆、近阅读、远距离阅读、开放获取等概念都在 h 子网中有所展示，这表明基于图书馆的研究在当前数字人文领域中得到了较多关注。此外，可以看到延伸出了一个比较小的网络分支，由人文学科、数字化、技术、创新性、数据、工具和信息所构成，表明了数字人文和人文学科的内在联系。在社交媒体分支上，被研究最多的平台仍然是 Twitter，此外，新产生了替代计量学（altmetrics）这一概念，在计量学领域中产生了影响。

4.3　知识吸收网络与知识扩散网络比较

将参考文献集、中心文献引文文本集和中心文献集的关键词以共现关系进行合并，表征数字人文核心关键词的知识吸收；将中心文献集、施引文献引文文本集和施引文献集的关键词以共现关系进行合并，表征数字人文核心关键词的知识扩散。

首先，比较分析知识吸收网络和知识扩散网络的数量特征。表 1 是知识吸收网络和知识扩散网络在数量特征上的对比表，包括整个网络特征及其 h 子网的特征。

表 1　知识吸收和知识扩散网络的数量特征比较

网络	网络节点数	网络联系数	网络密度	h 强度	h 子网节点数	h 子网密度
知识吸收	2 968	13 069	0.003	48	54	0.034
知识扩散	3 790	60 366	0.008	42	45	0.051

从表 1 中可以看到，知识扩散网络的整体规模大于知识吸收网络，包括整体网络的节点数、联系数和网络密度。这表明数字人文知识在扩散中分布的知识点更广泛，同时知识点之间整体的关联性更强。在 h 子网中，知识吸收网络中具有高联系强度的知识节点更多。

基于 h 子网从内容上比较知识吸收和知识扩散的核心网络。图 5 是知识吸收网络和知识扩散网络 h 子网中的节点集合图。从图 5 可以看到，两个集合中的节点具有较大交集。这些知识是数字人文领域知识吸收与扩散过程中的关键知识，除了数字人文概念之外，人文计算、地理信息系统相关知识与数字人文知识紧密相关。除此还有图书馆类知识，包括数字图书馆、档案、远距离阅读；学科类知识，包括史学、数字史学、人文学科、教育学等；数据数字化类知识，包括数字化、数据可视化、大数据等；还有文化遗产、Twitter 等社交媒体是关注的研究对象。在其中，有关本体、关联数据、语料库语言学、文本挖掘等相关技术是数字人文领域中的重点技术。

不同的是，在数字人文领域知识吸收中，更关注地图技术、基于树的主题模型、隐性信息等；而在数字人文领域知识扩散中，更关注信息基础设置、信息素养、嵌入式图书馆、替代计量学、开放获取、近距离阅读、持续性等。

5　基于引文上下文的数字人文知识引证关键词网络

5.1　高被引中心文献与参考文献的引证关键词网络

在中心文献对参考文献进行引用的过程中，构成了引文文本关键词对参考文献原关键词的引用关系，即若中心文献 A 引用了参考文献 B（原关键词：b1、b2、

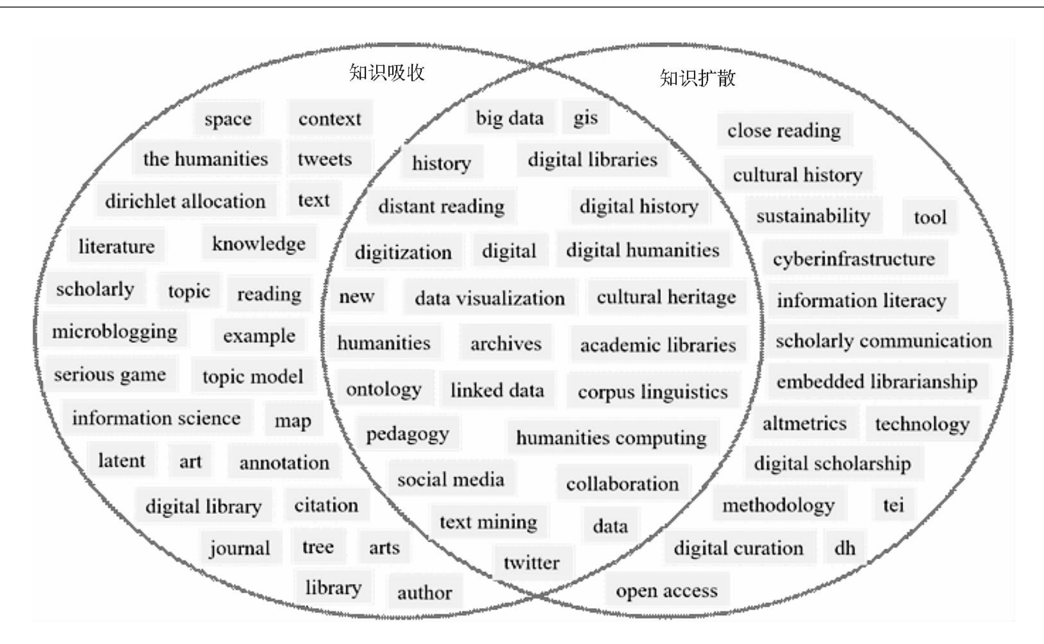


图 5　知识吸收与扩散 h 子网知识节点集合

b3),在中心文献 A 原文中识别出对应的引文文本 C(引文文本关键词:c1、c2、c3),则词 c1、c2、c3 对词 b1、b2、b3 分别构成引用关系,产生共 9 对引用关系对,即(c1,b1)(c1,b2)(c1,b3)(c2,b1)(c2,b2)(c2,b3)(c3,b1)(c3,b2)(c3,b3)。分别对引用关系词对进行统计,并对相同引用关系词对进行合并计数,其数值作为引用关系对之间的关系强度值。

基于词节点与节点之间的关系,形成高被引中心文献与其参考文献之间的引用网络。根据可视化情况,筛选高联系强度的节点及其关系进行展示,如图 6 所示:

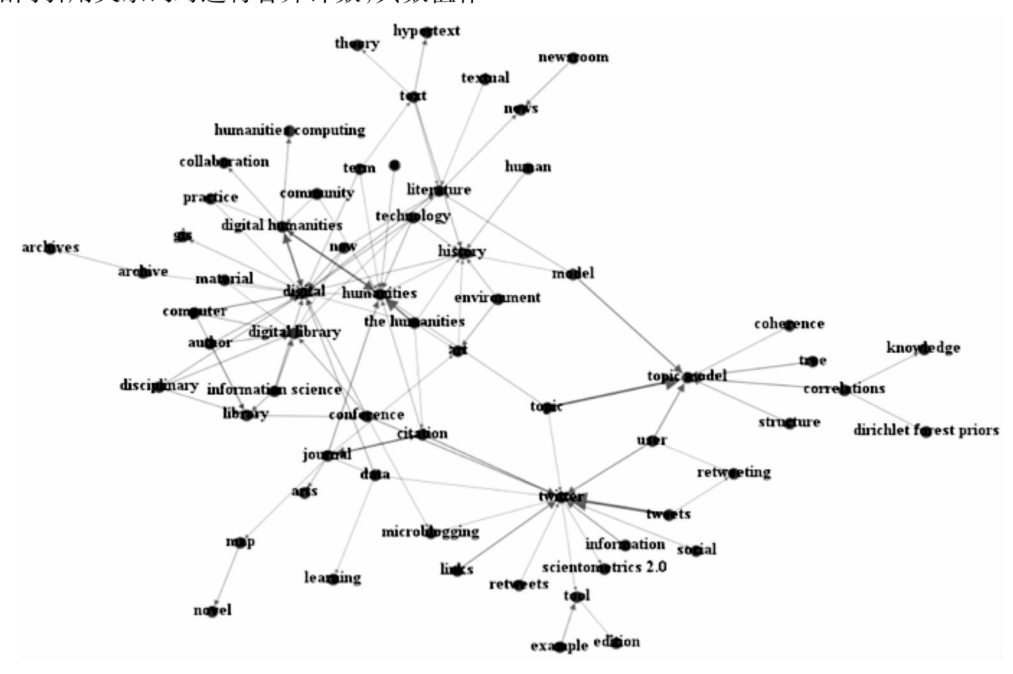


图 6　高被引中心文献与参考文献间的关键词引用网络核心

可以从图 6 中看到,Twitter、topic model、humanities 以及 digital 是网络中的核心节点,表明这几类是数字人文高被引知识形成过程中的重要信息。即在方向上,仍沿着数字人文的路径对人文学科进行数字化;在方法上,主题识别是研究过程中的重要手段;在载体上,Twitter 是国外数字人文研究的重要平台。

在人文学科与数字化研究的类群中,可以看到数字图书馆、社区、计算机、档案、GIS、期刊等都是数字人

文前期的重点研究对象。在主题识别的类群中,可以看到有关结构、"树"的算法等是研究方法中的关注点。在 Twitter 的类群中,可以看到有关用户、工具、社交等信息,表征着数字人文对社交平台研究的关注;可以明显看到 digital humanities 对 humanities computing 的引用,人文计算作为数字人文的前身,对数字人文的前期发展具有重要影响,以及历史、文学两个传统学科类别从一开始就是数字人文研究的重点学科领域。

除此之外,参考文献集的关键词补充了引文文本关键词后,可以揭示一些潜在的论文细节。例如 Twitter 和 topic model 是参考文献的关键词,在加入引文文本关键词后,可以看到其隐藏的路径关联 user,还可以看到与之连接的 retweeting,进而对研究细节进一步验证——对于社交平台的数字人文研究往往以用户为桥梁,主要分为两派:一是对内容进行主题分析;二是对转发行为进行关系研究。这是书目关键词网络难以挖掘的技术细节。

基于引文文本,将高被引中心文献的施引情况以关键词粒度进行可视化,可以探索出高被引中心文献中知识的形成过程,了解数字人文核心知识的来源,有利于从根源上了解数字人文,从而更好地把握数字人文的未来发展。

5.2 中心文献与施引文献的引证关键词网络

本研究从中心文献中共提取出 1 220 个关键词,实际是由 757 个词/词组所组成;从施引文本关键词中共提取出了 1 220 个关键词,实际是由 383 个词/词组所组成,其中有 113 个词/词组也属于被引文献关键词。被引词和施引词之间共产生了 5 508 对引用关系对。

首先,从被引词和施引词的实际词数组成可以看到,主要知识点从 757 个词/词组传递到 383 个词/词组,知识在引用过程中变得更集聚,领域研究关注点更加突出。表 2 是关系强度大于 30 的引用关系对,共有 15 组。

在表 2 中的被引词方面,digital humanities 出现的频数相对较高,这说明在显性信息中,明确围绕数字人文概念进行的研究能更多地引起学者们的共鸣,一方面关注数据、数字化深化,包括其中的技术手段;另一方面关注其研究内容,包括历史学研究或发展历史的研究。在表 2 中的施引词方面,Twitter 出现的频数是最高的,这表明从受众角度,数字人文领域已有研究传递出较多关于 Twitter 的有价值的信息。

表 2 关系强度大于 30 的引用词对

排名	被引词	施引词	关系强度	同词引用
1	digital humanities	data	70	
2	Twitter	Twitter	62	√
3	digital humanities	digital humanities	54	√
4	altmetrics	Twitter	47	
5	disciplinary differences	Twitter	47	
6	scholarly communication	Twitter	47	
7	webometrics	Twitter	47	
8	digital humanities	digital	46	
9	digital humanities	history	42	
10	social networks	Twitter	39	
11	humanities	digital humanities	36	
12	conferences	Twitter	36	
13	digital communication systems	Twitter	36	
14	user studies	Twitter	36	
15	digital humanities	technology	33	

在隐性信息中,Twitter 的高显示度表明了研究中对社交媒体这一对象的关注在不断提升,包括替代计量学、网络计量学等一系列由互联网进步所兴起的新兴计量指标,表明了数字人文研究在互联网领域的关注。

在引用过程中,还可以看到领域内对学科差异性、学术交流的重视。数字人文是一门交叉学科,整合学科差异、探索有效的跨学科交流合作模式能够有力推动数字人文的学术研究进展。

此外,用户研究、数字通讯系统、社会网络、相关会议等主题方向也得了学者们较多的关注。

在引用关系词对中,有一部分引用关系的被引词和施引词是同一个词/词组,表明这些词所代表的知识更多的是继承与深化,表 2 中就有部分关系词对是同词引用。在 5 508 对引用关系对中,有 61 对引用关系是同词引用。虽然在 5 508 对关系对中所占比例并不大,但从施引词的实际个数来看,384 个引用词中有 61 个词是同词引用。表明在引用过程中,至少有约 16% 的知识保持着同词传播与传承。表 3 是关系强度排名前 10 的同词引用关系对,关系强度最高的两对关系分别是词 Twitter 和 digital humanities。这表明在宏观层面上,围绕数字人文这一概念进行研究的知识多数也会以这一概念继续传播;在微观层面上,针对 Twitter 这一社交媒体进行的研究在被引用过程中,也会将 Twitter 作为一个特殊的研究对象进行传播。

其他高频同引关系对包括:新闻学、计算新闻学、档案、期刊等媒体相关的对象,技术、地理信息系统、数据等技术相关的对象,学术交流、引用等行为相关的对

表 3　同词引用的关系对 top10

排名	被引词	施引词	关系强度
1	Twitter	Twitter	62
2	digital humanities	digital humanities	54
3	journalism	journalism	21
4	computational journalism	computational journalism	19
5	technology	technology	16
6	scholarly communication	scholarly communication	15
7	archive	archive	12
8	journal	journal	11
9	data	data	10
10	citation	citation	9
10	GIS	GIS	9

象,在进行知识传播时,上述关键词是中心文献与施引文献间继承与深化的主动脉。

下文基于被引词与施引词之间的引用关系,构建数字人文研究领域的引用网络。图 7 是引文文本关键词与对应被引文献关键词所构建的引文网络,图中词节点大小代表词的中心度大小,箭头由施引词指向被引词。

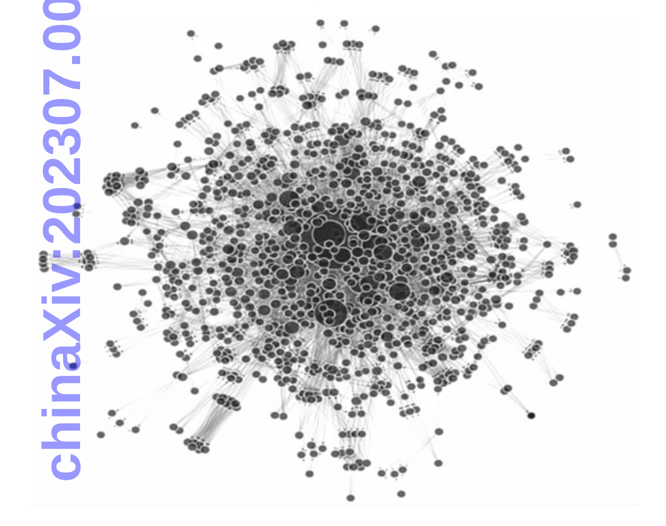


图 7　数字人文研究领域引用网络

从引用网络的图 7 中,能看到知识信息主要是从中心关键节点向外发散传播。中心节点间联系密切,反映了数字人文研究的热点之间知识结构较融合;外围节点间主要是“各成一派”,反映了数字人文在跨学科研究的背景下,研究方向具有差异性,许多研究知识之间保持独立性。

节点的中心度能反映节点在网络中所处的地位及权利影响,中心度高的在节点中处于核心地位,影响力大;反之,中心度低的在节点中处于边缘地位。网络中一个节点的点度中心度,可以用网络中与该点有直接联系的点的数目来衡量。在该数字人文研究领域的引用网络中,节点之间的平均度为 3.2,表明每一个词节点平均与其它 3 个词节点之间具有引用关系;节点之间的平均加权度为 7.2,表明每一个节点平均与其他节点具有共 7 次引用联系。

接下来通过限制网络中联系边的权重阈值来探索数字人文研究领域的核心引用网络。图 8 是通过限制节点间联系权重值不低于 20 所形成的引用网络,形成了两簇独立的引用网络,箭头由施引词指向被引词。

在左边的引用网络中,关键节点是 journalism(新闻)和 digital humanities(数字人文),连接了两边的引用网络。在 journalism 节点上,主要是其对其它知识的引用,表明在当前的研究中,新闻学方面的研究是一大热点,并且其综合了数据、技术、文化,在计算新闻、新制度主义、民族志、政治经济、新闻社会学等方面进行了深入的研究。在 digital humanities 节点上,主要是其它知识对其的引用,表明当前不少研究围绕并明确数字人文的概念进行探索,包括继续在数据信息、数字化、可视化、技术、史学等研究方向上进行深化。

在右边的网络中,相关研究主要围绕 Twitter 这一主题进行,并且主要是由 Twitter 从其他关键词方向获取信息进行整合研究,主要包括社交网络、用户研究、替代计量学、网络计量学、数字通信系统、会议等。而学科差异方向的研究会从 Twitter 研究中引用知识,认为其作为网络社交平台的代表,在探索新的学术交流方式、新的计量手段上具有重要的参照价值。

在有向网络中,从某个节点发出的定向连接的数量称为该接点的出度,指向该节点也终止于该节点的连接的数量称为该节点的入度。在数字人文研究的引

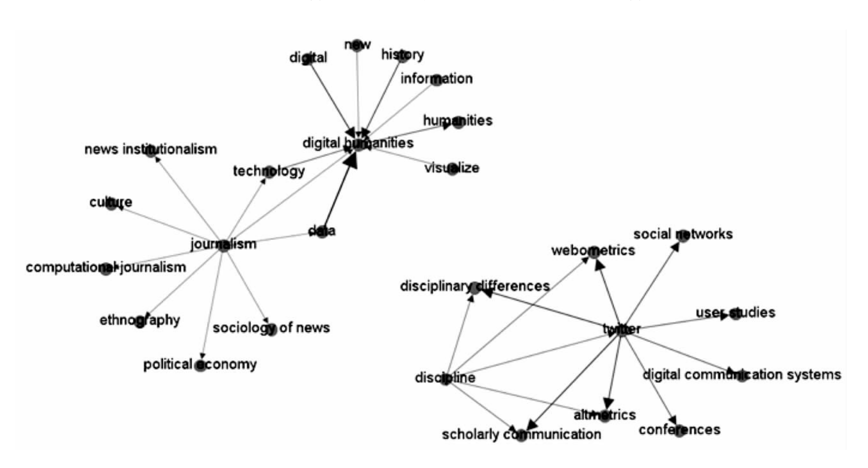


图 8　数字人文研究领域核心引用网络

用网络中，节点的入度表示被引关键词在被引用过程中的热度，也代表了数字人文研究传播出的主要知识；节点的出度表示在施引过程中学者们的关注点。表 4 是引用网络中入度最高的 10 个关键词和出度最高的 10 个关键词。

表 4 引用网络节点中心度

排名	关键词	入度（被引）	排名	关键词	出度（施引）
1	digital humanities	292	1	data	202
2	humanities	70	2	digital	167
3	history	50	3	digital humanities	163
4	archives	47	4	application	99
5	text mining	35	5	difference	96
6	academic libraries	32	6	technology	84
7	digital history	30	7	information	84
8	technology	27	8	new	82
9	social media	27	9	field	81
10	digital libraries	26	10	history	76

从节点入度排名可以看出，数字人文的概念、史学和数字史学、图书馆和档案、技术（包括文本挖掘）是传播最广的知识要素，是学者们当前最关注的数字人文研究领域的研究点。从出度排名可以看出，在引用过程中，学者们同样关注数字人文的概念、史学与技术，但更注重数字、数字化与应用，注重新的方法或手段所带来的差异性。以上对知识节点出入度的测量只考量了节点词与其他词的联系个数，测量了传播的广度，没有考虑节点词之间本身联系所具有的权重值。在构建的引用网络中，词与词之间的引用关系强度代表了原被引文献关键词在被引过程中，引文文本关键词对其引用的次数，本身反映了施引过程中的词间紧密程度。因此，综合词节点与其他各节点之间的联系及权重，得到各词节点的加权入度和加权出度，排名前 10 的词节点如表 5 所示：

表 5 引用网络节点加权中心度

排名	关键词	加权入度（被引）	排名	关键词	加权出度（施引）
1	digital humanities	1 091	1	data	648
2	humanities	278	2	Twitter	472
3	Twitter	169	3	digital humanities	467
4	technology	142	4	digital	422
5	scholarly communication	131	5	technology	330
6	culture	128	6	history	259
7	altmetrics	120	7	visualize	226
8	disciplinary differences	120	8	difference	224
9	webometrics	120	9	information	218
10	data	114	10	journalism	210

加入权重之后，更能反映出被引文献关键词传播的深度。从表 5 可以看到，文化、数据、Twitter 等对象的研究得以传承深化；而替代计量学、网络计量学等基于互联网的新兴词汇虽然在连接节点的广度上并不突出，但其加权入度值较高，可以看到其在数字人文研究领域具有小范围的深度传播。同时，学科的学术交流与学科差异也受到了学者们的重视。在加权出度方面，新增了对新闻、可视化两方面研究的关注。

6 结语

本文基于共现网络与引证关键词网络的方法，从知识受众的角度，建立了数字人文研究领域的新型知识图谱。在方法层面上，从引文上下文文本的特殊性出发，通过提取引文上下文并进行关键词识别，构建了引文文本关键词的共现网络。在此基础上，将引文文本关键词与被引文献的关键词进行连接，构建出新型引用网络，反映出知识与关注点在引用过程中的流动。在领域层面上，数字人文经过前期数年的知识积累，在近几年获得了广泛的关注。作为一种革命性的研究思想和有效的数字化工具，数字人文深入到各人文、艺术、社会科学以及地理、医学等众多学科方向，积累了较丰富的文献。对其领域发展进行梳理与总结，深入分析数字人文研究的热点，明确数字人文研究的演化路径和发展趋势，有利于了解数字人文研究的当前进展，为数字人文领域未来研究提供借鉴和参考，加快实践探索。

参考文献：

[1] 焦晓静，王兰成. 知识图谱的概念辨析与学科定位研究[J]. 图书情报工作，2015，59(15)：5－11.
[2] 李启虎，尹力，张全. 信息时代的人文计算[J]. 科学，2015，67(1)：35－39，4.
[3] 陈悦，刘则渊. 悄然兴起的科学知识图谱[J]. 科学学研究，2005，23(2)：149－154.
[4] 张斌，马费成. 科学知识网络中的链路预测研究述评[J]. 中国图书馆学报，2015，41(3)：99－113.
[5] 郑彦宁，许晓阳，刘志辉. 基于关键词共现的研究前沿识别方法研究[J]. 图书情报工作，2016，60(4)：1－8.
[6] 吴晓秋，吕娜. 基于关键词共现频率的热点分析方法研究[J]. 情报理论与实践，2012，35(8)：115－119.
[7] 宋艳辉，武夷山. 作者文献耦合分析与作者关键词耦合分析比较研究：Scientometrics 实证分析[J]. 中国图书馆学报，2014，40(1)：25－38.
[8] 张玲玲，张宇娥，杜丽. 国家社科基金项目成果视角下图情领域知识扩散研究[J]. 图书馆工作与研究，2017，1(10)：60－66.

[9] 罗双玲，张文琪，夏昊翔. 基于半积累引文网络社区发现的学科领域主题演化分析——以"合作演化"领域为例[J]. 情报学报，2017,36(1):100-110.

[10] 刘洋，崔雷. 引文上下文在文献内容分析中的信息价值研究[J]. 图书情报工作，2014,58(6):101-104.

[11] LIU Y，RAFOLS I，ROUSSEAU R. A framework for knowledge integration and diffusion[J]. Journal of documentation，2012,68(1):31-44.

[12] BORNMANN L，HAUNSCHILD R，HUG S E. Visualizing the context of citations referencing papers published by Eugene Garfield: a new type of keyword co-occurrence analysis[J]. Scientometrics，2018,114(2):427-437.

[13] CHEN C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology，2006,57(3):359-377.

[14] 肖明，陈嘉勇，李国俊. 基于CiteSpace研究科学知识图谱的可视化分析[J]. 图书情报工作，2011,55(6):91-95.

[15] 范云满，马建霞，曾苏. 基于知识图谱的领域新兴主题研究现状分析[J]. 情报杂志，2013(9):88-94.

[16] PHELPS C，HEIDL R，WADHWA A. Knowledge，networks，and knowledge networks: a review and research agenda[J]. Journal of management,2012，38(4):1115-1166.

[17] SCHREIBMAN S，SIEMENS R，UNSWORTH J. A companion to digital humanities[M]. New Jersey: John Wiley & Sons，2008:20-39.

[18] COOPER D，GREGORY I N. Mapping the English Lake District: a literary GIS[J]. Transactions of the institute of british geographers，2015,36(1):89-108.

[19] HINRICHS U，FORLINI S，MOYNIHAN B. Speculative practices: utilizing InfoVis to explore untapped literary collections[J]. IEEE transactions on visualization & computer graphics，2016,22(1):429-438.

[20] WONG S H R. Digital humanities: what can libraries offer? [J]. Portal: libraries and the academy，2016,16(4): 669-690.

[21] 柯平，宫平. 数字人文研究演化路径与热点领域分析[J]. 中国图书馆学报，2016,42(6):13-30.

[22] 高胜寒,赵宇翔,朱庆华. 国内外数字人文领域研究进展分析[J]. 图书馆杂志,2016,35(10):9-18.

[23] WANG Q. Distribution features and intellectual structures of digital humanities: a bibliometric analysis[J]. Journal of documentation，2018,74(1):223-246.

[24] ALJABER B，STOKES N，BAILEY J，et al. Document clustering of scientific texts using citation contexts[J]. Information retrieval，2010,13(2):101-131.

[25] BRADSHAW S. Reference directed indexing: redeeming relevance for subject search in citation indexes[M]// Research and advanced technology for digital libraries. Berlin: Springer，2003:499-510.

作者贡献说明：
许鑫:负责论文选题、研究方案设计；
陈路遥:负责论文资料收集、数据分析和初稿撰写；
杨佳颖:负责后续修改,形成定稿。

Knowledge Network in the Digital Humanities Domain
——Based on the Analysis of Keywords and Citation Contexts in WoS

Xu Xin　Chen Luyao　Yang Jiaying
Department of Information Management，Faculty of Economics and Management，
East China Normal University，Shanghai 200241

Abstract: [**Purpose/significance**] Citation shows the link between citing articles and cited ones，which reflects the reference and affirmation of successors. Derived from the traditional keyword network，this paper has innovated keyword network based on citation context. The constructed knowledge map can not only reveal the deep information of the literature，but also reflect the knowledge-based process by which the readers actively select and utilize the literature. [**Method/process**] In this paper，digital humanity was established as the research field. Three literature sets and two citation text sets were collected to build two non-direction keyword-sharing networks and two directed keyword networks based on citation. The co-occurrence network showed the absorption and diffusion of the knowledge of digital humanities while the citation keyword network illustrated the formation and transformation of digital humanities. [**Result/conclusion**] After visualization of the constructed network，the research obtained core domain and core technology of digital humanities，which provided reference for the future research of digital humanities from the perspective of information recipients.

Keywords: mapping knowledge domain　digital humanities　citation context　keyword network　visualization